Statistica Sinica Preprint No: SS-2020-0440	
Title	Adaptive Randomization via Mahalanobis Distance
Manuscript ID	SS-2020-0440
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0440
Complete List of Authors	Yichen Qin, Yang Li, Wei Ma, Haoyu Yang and Feifang Hu
Corresponding Authors	Feifang Hu
E-mails	feifang@gwu.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

Adaptive Randomization via Mahalanobis Distance

Yichen Qin[1], Yang Li[2], Wei Ma[2], Haoyu Yang[2], and Feifang Hu[3]

[1] *University of Cincinnati*

[2] *Renmin University of China*

[3] *George Washington University*

Abstract:

In comparative studies, researchers often seek an optimal covariate balance. However, chance imbalance still exists in randomized experiments, and becomes more serious as the number of covariates increases. To address this issue, we introduce a new randomization procedure, called adaptive randomization via the Mahalanobis distance (ARM). The proposed method allocates units sequentially and adaptively, using information on the current level of imbalance and the incoming unit's covariate. Theoretical results and numerical comparison show that with a large number of covariates or a large number of units, the proposed method shows substantial advantages over traditional methods in terms of the covariate balance, estimation accuracy, hypothesis testing power, and computational time. The proposed method attains the optimal covariate balance, in the sense that the estimated treatment effect attains its minimum variance asymptotically, and can be applied in both causal inference and clinical trials. Lastly, numerical stud-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

ies and a real–data analysis provide further evidence of the advantages of the proposed method.

Key words and phrases: covariate balance, clinical trial; treatment effect estimation.

1. Introduction

Randomization is the foundation of evaluating a treatment effect. However, traditional randomization methods often generate unsatisfactory configurations, with unbalanced prognostic covariates. "*Most of experimenters on carrying out a random assignment of plots will be shocked to find out how far from equally the plots distribute themselves*"(Fisher, 1926). Balanced covariates offer three main advantages (Hu et al., 2014). First, it improves the efficiency of a treatment effect estimation. Second, it increases the interpretability of the estimated treatment effect by making the units in the treatment groups more comparable, thereby enhancing the credibility of the analysis. Third, it makes the analysis more robust against model misspecification.

if significant covariate imbalance exists in clinical studies and causal inference, the subsequent inference on the treatment effect often needs to be adjusted. Some ex-post adjustments, such as regression (Freedman, 2008)

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

and subsample selection using matching or trimming based on propensity scores (Imbens and Rubin, 2015), can cope with such imbalance, but are much less efficient than achieving an ex-ante balance (Bruhn and McKenzie, 2008). In addition, these adjustments often rely on at least a nearly correct model, which can be difficult to test (Cochran, 1965; Cochran and Rubin, 1973). Rubin (2008) that the design phase of an experiment is particularly important, because in the analysis stage, the researcher may bias the results (Lock, 2011; Imbens and Rubin, 2015).

Furthermore, the effects of a covariate imbalance become worse as the number of covariates p and the sample size n become large, as is almost always the case in big data. Although the difference between the covariate means across treatment groups becomes smaller as n increases, the confidence intervals become more sensitive to small differences in the outcome variables, which can be affected by imbalances on the covariates (Morgan and Rubin, 2012).

In the context of causal inference, Morgan and Rubin (2012) propose rerandomization (RR). They propose repeatedly randomizing the units into treatment groups using complete randomization (CR), until a balance criterion, namely, the Mahalanobis distance $M(n)$ between the sample means

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

across different treatment groups, is below a threshold $a > 0$:

$$M(n) = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T [\mathrm{cov}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)]^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \propto n(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2),$$

where $\bar{\boldsymbol{x}}_1 \in \mathbb{R}^p$ and $\bar{\boldsymbol{x}}_2 \in \mathbb{R}^p$ are the sample means for two treatment groups, and $\boldsymbol{\Sigma} = \mathrm{cov}(\boldsymbol{x}) \in \mathbb{R}^{p \times p}$ is the covariance matrix of the covariate. Morgan and Rubin (2012) also assume fixed equal numbers of units in the two treatment groups, and demonstrate various desirable properties. RR works well in the case of a few covariates. However, as the number of covariates increases, the probability of acceptance, $P_a = P(M(n) < a)$, decreases drastically, causing the RR procedure to remain in a loop for a long time. To alleviate the computational burden, one can increase a, which may lead to a covariate imbalance.

In clinical trials, to balance important covariates, most existing methods focus only on discrete covariates. These methods include stratified permuted block randomization (SPBR), the stratified biased coin design (SBCD) (Shao et al., 2010), minimization methods (Taves, 1974; Pocock and Simon, 1975; Hu and Hu, 2012), and the covariate-adaptive biased coin design (CA-BCD) (Antognini and Zagoraiou, 2011). Discretizing continuous covariates is often less efficient and changes the nature of the covariates. A variety of methods for balancing continuous covariates have been proposed, including optimum biased coin designs (DA-BCD) (Atkinson, 1982)

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

and methods based on ranks (Ciolino et al., 2011; Hoehler, 1987; Stigsby and Taves, 2010), p-value (Frane, 1998), the Kullback–Leibler divergence (KLD), an empirical cumulative distribution (Lin and Su, 2012), and the kernel density (Ma and Hu, 2013). However, the performance of these procedures is usually evaluated by simulation, and few studies examine their theoretical properties.

Here, we propose an approach called adaptive randomization via the Mahalanobis distance (ARM). The ARM approach generates a more balanced treatment allocation, and thus improves the subsequent estimation and inference in causal inference and clinical trial settings. Unlike RR and CR, in which all units are allocated independently, we allocate units adaptively and sequentially by assigning one randomly chosen pair of units at a time. For each pair of units, we avoid incidental covariate imbalance by using their covariate information and the existing level of imbalance of the already allocated units to adjust the probability with which the pair is allocated to a treatment group. In this way, we produce a much more balanced allocation of units. We investigate the properties of the proposed procedure both theoretically and numerically.

The proposed method offers several advantages. First, when we have a large number of covariates or a large number of units, the proposed method

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

exhibits superior performance, with a more balanced randomization and less computational time. Second, the proposed procedure attains the optimal covariate balance, in the sense that the estimated treatment effect under the proposed method attains its minimum variance asymptotically. Third, in addition to the optimal estimation precision, the proposed procedure is the most powerful of several tests for the treatment effect. Fourth, the proposed procedure is designed to directly randomize units with continuous and discrete covariates. Therefore, the ARM procedure can be applied to balance many important covariates in comparative studies.

The remainder of this paper is organized as follows. We introduce the proposed ARM method and investigate its theoretical properties in Section 2. We demonstrate its advantages in terms of treatment effect estimation and hypothesis testing in Section 3. In Sections 4 and 5, we use simulations and a real–data analysis, respectively, to demonstrate the superior performance of the proposed method. Section 6 concludes the paper. All proofs are related to the online supplementary material.

2. The ARM method

Suppose that n units (patients) are assigned to two treatment groups. Let T_i be the assignment of the ith unit, that is, $T_i = 1$ for treatment 1, and

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

$T_i = 0$ for treatment 2. Consider p continuous covariates for each unit. Let $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T \in \mathbb{R}^p$ represent the covariates of the ith unit. For simplicity, we first assume a causal inference setting in which all units are available for assignment at the beginning of the randomization. Later, we explain how to adapt this for a clinical trial setting in which the units enroll in the study sequentially. The ARM method is as follows:

(1) Arrange all n units randomly into a sequence $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$.

(2) Assign the first two units with $T_1 = 1$ and $T_2 = 0$.

(3) Suppose that $2i$ units have been assigned to treatment groups, for the $(2i+1)$th and $(2i+2)$th units:

(3a) If the $(2i+1)$th unit is assigned to treatment 1 and the $(2i+2)$th unit is assigned to treatment 2, then we can calculate the "potential" Mahalanobis distance between the updated treatment groups with $2i+2$ units, that is, $M_1(2i+2) = (\bar{\boldsymbol{x}}_{1,2i+2} - \bar{\boldsymbol{x}}_{2,2i+2})^T \widehat{\boldsymbol{\Sigma}}^{-1} (\bar{\boldsymbol{x}}_{1,2i+2} - \bar{\boldsymbol{x}}_{2,2i+2})$, where $\widehat{\boldsymbol{\Sigma}}$ is the sample covariance matrix, and $\bar{\boldsymbol{x}}_{1,2i+2}$ and $\bar{\boldsymbol{x}}_{2,2i+2}$ are the updated covariate means of the treatment groups.

(3b) Similarly, if the $(2i+1)$th unit is assigned to treatment 2 and the $(2i+2)$th unit is assigned to treatment 1, then we can calculate

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

the other "potential" Mahalanobis distance, $M_2(2i+2)$.

(4) Assign the $(2i+1)$th unit to treatment groups according to the following probabilities:

$$P(T_{2i+1}=1|\boldsymbol{x}_{2i},...,\boldsymbol{x}_1,T_{2i},...,T_1)=\begin{cases} q & \text{if } M_1(2i+2)<M_2(2i+2), \\ 1-q & \text{if } M_1(2i+2)>M_2(2i+2), \\ 0.5 & \text{if } M_1(2i+2)=M_2(2i+2), \end{cases}$$

where $0.5<q<1$, and assign $T_{2i+2}=1-T_{2i+1}$ to maintain equal proportions.

(5) Repeat the steps 3 and 4 until all units are assigned. If n is odd, assign the last unit to two treatments with equal probabilities.

We choose the Mahalanobis distance as the covariate imbalance measure for several reasons. First, the Mahalanobis distance has produced good results and is used frequently in the literature (Morgan and Rubin, 2012, 2015; Li et al., 2018; Zhou et al., 2018; Li and Ding, 2020). Furthermore, it is an affinely invariant imbalance measure, which is appealing for multivariate data. Second, we can obtain desirable properties by using the Mahalanobis distance, such as the optimal asymptotic variance for the treatment effect estimation. Third, in real–data analyses, covariates often have different variances, ranges, and metrics (e.g., inches and pounds). Given limited computational power, minimizing the Mahalanobis distance evenly improves the balance of all covariates, because it considers the covariate balance relative to their variances. During the randomization stage,

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

it is difficult to know which of the covariates contribute most to the balance, because the outcome variables are not observed. Therefore, minimizing the Mahalanobis distance is safer, because it improves the balance of all covariates equally.

Note that we are not establishing the Mahalanobis distance as the only choice of imbalance measure. Other choices, such as the L2–norm of the imbalance vector, are also available, and focus on different aspects of the covariate imbalance. When the covariance matrix is the identity matrix, the Mahalanobis distance becomes the L2–norm of the imbalance vector. The aforementioned algorithm also works with difference imbalance measures.

In steps (3a) and (3b), the covariance matrix in the Mahalanobis distance is replaced with the sample covariance matrix $\widehat{\mathbf{\Sigma}}$. In a causal inference setting, the sample covariance matrix is based on all units, $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}_n$, and remains the same throughout the randomization. In a clinical trial setting, the sample covariance estimate is updated using the first $2i + 2$ units during each iteration, $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}_{2i+2}$. The proposed procedure can adapt easily to both settings.

Here, we allocate a pair of units at a time so that the sample sizes across the treatment groups remain equal, $\sum_{i=1}^{n} T_i = \sum_{i=1}^{n}(1 - T_i)$. However, depending on the speed of the patient recruitment process, we can allocate

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

unita one at a time, but at the cost of different treatment group sample sizes. In practice, this modified version of the ARM approach also performs well, and we denote it as mARM. The algorithm is provided in the Supplementary Material.

The value of q is set to 0.75 throughout this article. Different values of q do not affect our theoretical results. For a further discussion of q, please see Hu and Hu (2012).

The proposed method is designed for directly randomizing units with continuous covariates. In the literature, continuous covariates are usually discretized on order to be included in the balancing procedures. However, breaking a continuous covariate into subcategories means increased effort and a loss of information, as pointed by Scott et al. (2002). Ciolino et al. (2011) further note that the lack of publicity about practical methods for continuous covariate balancing and the lack of knowledge about the cost of failing to balance continuous covariates results in continuous covariates being excluded from the randomization plan in clinical trials. Therefore, the proposed method also contributes to the literature in this regard. Note that, the proposed procedure works well for large p and n, whereas most existing methods work only for small p.

We now study the asymptotic properties of the Mahalanobis distance

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

under the proposed method.

Theorem 1. *Under the proposed procedure, suppose that the covariate* $\boldsymbol{x}_i$*, for* $i = 1, ..., n$*, is independent and identically distributed (i.i.d) as a multivariate normal distribution with a zero mean and covariance matrix* $\boldsymbol{\Sigma}$*. Then we have* $M(n) = O_p(n^{-1})$*.*

Note that the Mahalanobis distance obtained using CR has a chi-squared distribution with p degrees of freedom, that is, $M_{\text{CR}}(n) \sim \chi^2_{df=p}$. The Mahalanobis distance obtained using RR has a conditional chi-squared distribution, that is, $M_{\text{RR}}(n) \sim \chi^2_{df=p} | \chi^2_{df=p} < a$. Hence, the proposed method outperforms RR and CR as the sample size increases, because its Mahalanobis distance converges to zero at a rate of $1/n$. That is, as more units are included, the better the covariate balance becomes. In addition, note that our theoretical results focus on the case of fixed p and diverging n.

Moreover, as the number of covariates p increases, the distribution of $M_{\text{CR}}(n)$ becomes flatter, which implies a poorer covariate balance. Consequently, RR has a lower probability of acceptance. Therefore, the advantage of the proposed method becomes more significant as p increases, because the $M(n)$ obtained using the proposed method converges to zero regardless of p.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

3. Treatment Effect Estimation and Hypothesis Testing

3.1 Framework

After the randomization, we estimate the treatment effect based on the outcome variable y_i obtained under treatment T_i, for $i = 1, ..., n$. A natural choice is the difference-in-means estimator,

$$\hat{\tau} = \frac{\sum_{i=1}^{n} T_i y_i}{\sum_{i=1}^{n} T_i} - \frac{\sum_{i=1}^{n} (1 - T_i) y_i}{\sum_{i=1}^{n} (1 - T_i)}. \tag{3.1}$$

However, $\hat{\tau}$ is sensitive to covariate imbalance. For example, if treatment 1 contains mostly males and treatment 2 contains mostly females, then $\hat{\tau}$ will not be able to exclude the gender effect.

To adjust for such an imbalance, we can use a linear regression to estimate the treatment effect. That is, conditional on the treatment assignment T_i, the outcome variable is assumed to follow

$$y_i = \mu_1 T_i + \mu_2 (1 - T_i) + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i, \tag{3.2}$$

where μ_1 and μ_2 are the main effects of treatments 1 and 2, respectively, and $\mu_1 - \mu_2 = \tau$ is the treatment effect. Furthermore, β_j represents the covariate effect, and ϵ_i is an i.i.d. random error with a zero mean and constant variance σ_ϵ^2, and is independent of the covariates.

Define $\boldsymbol{Y} = (y_1, ..., y_n)^T$, $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)^T = [x_{ij}]_{n \times p}$, $\boldsymbol{T} = (T_1, ..., T_n)^T$, $\mathbf{1} = (1, ..., 1)^T$, $\widetilde{\boldsymbol{X}} = [\boldsymbol{T}; \mathbf{1} - \boldsymbol{T}; \boldsymbol{X}]$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$, and $\boldsymbol{\beta}^* = (\mu_1, \mu_2, \boldsymbol{\beta}^T)^T$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

The ordinary least squares estimate of $\boldsymbol{\beta}^*$ is $\widehat{\boldsymbol{\beta}}^* = (\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}^T\boldsymbol{Y}$. Consider $\boldsymbol{L} = (1, -1, 0, ..., 0)^T$, a $(p+2)$-dimensional vector. We define the linear-regression-adjusted estimator under (3.2) as

$$\tilde{\tau} = \boldsymbol{L}^T\widehat{\boldsymbol{\beta}}^*,$$

which is adjusted for the covariate imbalance. Note that if $\widetilde{\boldsymbol{X}}$ does not include any covariates, that is, $\widetilde{\boldsymbol{X}} = [\boldsymbol{T}; \boldsymbol{1} - \boldsymbol{T}]$, then the working model becomes

$$y_i = \mu_1 T_i + \mu_2(1 - T_i) + \epsilon_i. \tag{3.3}$$

Hence, $\tilde{\tau}$ under (3.3) becomes $\hat{\tau}$, the difference-in-means estimator.

In addition to estimating treatment effects, hypothesis testing is another important part of comparative studies. To detect whether a treatment effect exists, we have the following hypothesis testing problem:

$$H_0 : \mu_1 - \mu_2 = 0 \text{ versus } H_1 : \mu_1 - \mu_2 \neq 0. \tag{3.4}$$

We can conduct the hypothesis tests under models (3.2) and (3.3), respectively, which essentially correspond to using $\tilde{\tau}$ and $\hat{\tau}$, respectively.

Under model (3.2), we define the test statistic S_{adj} as

$$S_{\text{adj}} = \frac{\boldsymbol{L}^T\widehat{\boldsymbol{\beta}}^*}{\sqrt{\hat{\sigma}_w^2\boldsymbol{L}^T\left(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}}\right)^{-1}\boldsymbol{L}}} = \frac{\tilde{\tau}}{\sqrt{\hat{\sigma}_w^2\boldsymbol{L}^T\left(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}}\right)^{-1}\boldsymbol{L}}},$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

where $\hat{\sigma}_w^2 = \|\boldsymbol{Y} - \widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}}^*\|^2/(n-p-2)$ is the estimate of σ_ϵ^2 under model (3.2). The test statistic S_{adj} is essentially a linear–regression–adjusted T-test for the treatment effect $\mu_1 - \mu_2$. Traditionally, the null hypothesis is rejected if $|S_{\text{adj}}| > z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal distribution, and α is the significance level.

Under model (3.3), we can simplify the test statistic by letting $\widetilde{\boldsymbol{X}} = [\boldsymbol{T}; \boldsymbol{1} - \boldsymbol{T}]$ and $\boldsymbol{L} = (1, -1)^T$. The test statistic becomes

$$S_{\text{unadj}} = \frac{\hat{\tau}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where $n_1 = \sum_{i=1}^n T_i$, $n_2 = n - n_1$, and $s^2 = (\sum_{i:T_i=1}(y_i - \bar{y}_1)^2 + \sum_{i:T_i=0}(y_i - \bar{y}_2)^2)/(n-2)$, where $\bar{y}_1$ and $\bar{y}_2$ are the respective sample means of the two treatment groups. Note that this is essentially a two–sample T–test statistic with equal variance, but it is not adjusted for the covariate imbalance. Again, the null hypothesis is rejected if $|S_{\text{unadj}}| > z_{1-\alpha/2}$.

Note that throughout this paper, the outcome variable genuinely follows model (3.2), where the covariates can directly affect the outcome variable. However, we often use the working model (3.3) for inference, for various practical reasons, such as the simplicity of the testing procedure and its robustness to model misspecification, even though some covariates are omitted in the working model (3.3).

In the next section, we examine the properties of estimating using by $\tilde{\tau}$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

and $\hat{\tau}$ and hypothesis testing using S_{adj} and S_{unadj} under the proposed and other randomization methods.

3.2 Theoretical Properties

Under the framework above, we can show that $\hat{\tau}$ under the ARM method achieves the optimal precision, and S_{unadj} under the ARM method achieves the highest power.

Theorem 2 (Optimal precision)**.** *Suppose that the outcome variable y_i follows the linear regression model in Equation* (3.2)*, and that we estimate the treatment effect under the proposed method and under CR; then, we have*

$$\sqrt{n}\big(\hat{\tau}_{\text{ARM}} - (\mu_1 - \mu_2)\big) \xrightarrow{D} N(0, V_1), \quad \sqrt{n}\big(\tilde{\tau}_{\text{ARM}} - (\mu_1 - \mu_2)\big) \xrightarrow{D} N(0, V_2),$$

$$\sqrt{n}\big(\tilde{\tau}_{\text{CR}} - (\mu_1 - \mu_2)\big) \xrightarrow{D} N(0, V_3), \quad \sqrt{n}\big(\hat{\tau}_{\text{CR}} - (\mu_1 - \mu_2)\big) \xrightarrow{D} N(0, V_4),$$

where $4\sigma_\epsilon^2 = V_1 = V_2 = V_3 < V_4$.

This theorem implies that under the proposed method, the precision of the difference-in-means estimator, $\hat{\tau}_{\text{ARM}}$, is the same as the precision of the linear-regression-adjusted estimator, $\tilde{\tau}_{\text{ARM}}$. This suggests that the proposed method can balance the covariates so well that, asymptotically, the regression adjustment is not needed, and $\hat{\tau}_{\text{ARM}}$ is just as good as $\tilde{\tau}_{\text{ARM}}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

In addition, under the linear regression assumption, we can show that

$$\hat{\tau} = \mu_1 - \mu_2 + \frac{2}{n}\left[\sum_{i=1}^{n}(2T_i - 1)\,\epsilon_i + \sum_{j=1}^{p}\beta_j\left(\sum_{i\in\{i:T_i=1\}} x_{ij} - \sum_{i\in\{i:T_i=0\}} x_{ij}\right)\right].$$

Therefore, the variance of $\hat{\tau}$ can be decomposed into two parts:

$$\mathrm{Var}(\hat{\tau}) = \mathrm{Var}\left(\frac{2}{n}\sum_{i=1}^{n}(2T_i - 1)\,\epsilon_i\right) + \mathrm{Var}\left(\frac{2}{n}\sum_{j=1}^{p}\beta_j\left(\sum_{i\in\{i:T_i=1\}} x_{i,j} - \sum_{i\in\{i:T_i=0\}} x_{i,j}\right)\right). \tag{3.5}$$

The first part of $\mathrm{Var}(\hat{\tau})$ is due to the random error, and the second is due to the covariate imbalance. Because random errors are inevitable, the minimum variance of $\hat{\tau}$ is $\mathrm{Var}\left(2\sum_{i=1}^{n}(2T_i - 1)\,\epsilon_i/n\right) = 4\sigma_\epsilon^2/n$. Theorem 2 shows that the precision of $\hat{\tau}_{\mathrm{ARM}}$ is exactly $4\sigma_\epsilon^2/n$. Therefore, we conclude that the difference-in-means estimator under the proposed method, $\hat{\tau}_{\mathrm{ARM}}$, attains the optimal precision, but that $\hat{\tau}_{\mathrm{CR}}$ cannot.

Theorem 2 further implies that the variance of the linear-regression-adjusted estimator $\tilde{\tau}$ is also $4\sigma_\epsilon^2/n$, regardless of the randomization method, as long as the linear model assumption is true. That is, the second term of (3.5) can be eliminated by using regressions. Therefore, $\tilde{\tau}$ is also considered asymptotically optimal, with the help of a linear regression.

Although $\tilde{\tau}$ and $\hat{\tau}_{\mathrm{ARM}}$ have the same precision, note that $\hat{\tau}_{\mathrm{ARM}}$ is a sample mean difference and is conceptually simple, whereas $\tilde{\tau}$ needs to estimate all regression coefficients $\boldsymbol{\beta}^*$ and requires linear regression assumptions. In some situations, $\tilde{\tau}$ is not preferred, such as in the case of model misspec-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

ification, ethical issues, ex-post adjustments, data privacy, and so on. Li and Ding (2020) examine a more general case in which the covariates in the design and analysis stages can take various relationships, and emphasize the importance of ex-ante balance. In contrast, we focus on the case in which the covariates in the analysis stage are subsets of those in the design stage, and provide similar conclusions. In the Supplementary Material, we provide additional numerical studies that compare the performance of $\hat{\tau}$ and $\tilde{\tau}$ under several model misspecification cases.

For comparison, we present the properties of $\hat{\tau}_{\mathrm{RR}}$ and $\tilde{\tau}_{\mathrm{RR}}$. Note that all properties are derived under the proposed framework, which differs from that of Morgan and Rubin (2012).

Corollary 1. *Under the same assumptions as in Theorem 2, suppose that we estimate the treatment effect under the RR; then, we have*

$$\sqrt{n}\big(\tilde{\tau}_{\mathrm{RR}} - (\mu_1 - \mu_2)\big) \xrightarrow{D} N(0, V_5), \quad \sqrt{n}\big(\hat{\tau}_{\mathrm{RR}} - (\mu_1 - \mu_2)\big) \xrightarrow{D} N(0, V_6),$$

where $4\sigma_\epsilon^2 = V_1 = V_2 = V_3 = V_5 < V_6 < V_4$.

This theorem shows that RR cannot achieve optimal precision in contrast to the proposed method. Furthermore, it cannot completely remove the covariate imbalance. In Table 1, we summarize the relationships between these estimators' asymptotic variances.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

Table 1: The relationship of asymptotic variances of different estimators. All results are derived under the proposed framework.

Randomized Covariates	Randomization Method	Estimators $\hat{\tau}$		$\tilde{\tau}$
	CR	Asym. Var.	$>$	Asym. Var.
		$\vee$		$\|\|$
$\boldsymbol{X}$	RR	Asym. Var.	$>$	Asym. Var.
		$\vee$		$\|\|$
	ARM	Asym. Var.	$=$	Asym. Var.

Next, we establish the properties of the hypothesis tests using S_{adj} and S_{unadj}. Note that because of the optimality of $\hat{\tau}_{\text{ARM}}$, S_{unadj} under ARM also attains the highest power among all tests. Theorem 3 follows from the work of Ma et al. (2019).

Theorem 3. *Under the proposed method, when testing the treatment effect using S_{unadj} (i.e., the two-sample T-test with equal variance), we have*

1. *Under $H_0 : \mu_1 - \mu_2 = 0$; then, $S_{\text{unadj}} \xrightarrow{d} N\left(0, \sigma_\epsilon^2/(\sigma_\epsilon^2 + \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta})\right)$.*

2. *Under $H_1 : \mu_1 - \mu_2 \neq 0$, where $\mu_1 - \mu_2 = \delta/\sqrt{n}$, for a fixed $\delta \neq 0$; then, $S_{\text{unadj}} \xrightarrow{d} N\left(\delta/\left(2\sqrt{\sigma_\epsilon^2 + \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}}\right), \sigma_\epsilon^2/(\sigma_\epsilon^2 + \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta})\right)$.*

Theorem 3 provides insights about the distribution of S_{unadj} under the ARM method. We often incorrectly assume a standard normal distribution for S_{unadj} under the null hypothesis. However, according to our results, under the ARM method, the null distribution of S_{unadj} is narrower than the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

standard normal distribution, because $\sigma_\epsilon^2/(\sigma_\epsilon^2 + \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}) < 1$. Therefore, the traditional testing procedure with the critical value $z_{1-\alpha/2}$ leads to a reduced type-I error.

To obtain the correct type I–error, we can adjust the testing procedure using the corrected critical value $z_{1-\alpha}^{\text{ARM}}$ based on part 1 of Theorem 3. This test also has an adjusted test has higher power than that of the traditional test. In practice, because the corrected critical value depends on the unknown parameters, we use the bootstrap method or directly estimate the parameters to obtain the critical value.

Similarly, we establish the asymptotic distribution of S_{unadj} under CR and RR as follows (Ma et al., 2019).

Corollary 2. *Under CR, when testing the treatment effect using* S_{unadj} *(i.e., the two–sample T-test with equal variance), we have*

1. *Under* $H_0 : \mu_1 - \mu_2 = 0$*; then,* $S_{\text{unadj}} \xrightarrow{d} N(0, 1)$.

2. *Under* $H_1 : \mu_1 - \mu_2 \neq 0$*, where* $\mu_1 - \mu_2 = \delta/\sqrt{n}$*, for a fixed* $\delta \neq 0$*; then,* $S_{\text{unadj}} \xrightarrow{d} N\left(\delta/\left(2\sqrt{\sigma_\epsilon^2 + \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}}\right), 1\right)$.

Corollary 3. *Under RR, when testing the treatment effect using* S_{unadj} *(i.e., the two–sample T-test with equal variance), we have*

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

1. *Under* $H_0 : \mu_1 - \mu_2 = 0$*; then,* $S_{\text{unadj}} \xrightarrow{d} (\sigma_\epsilon Z + \boldsymbol{\beta}^T \boldsymbol{\xi}) / \left(\sqrt{\sigma_\epsilon^2 + \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta}} \right)$*, where* Z *is a standard normal random variable and* $\boldsymbol{\xi} = \boldsymbol{\Sigma}^{1/2} \boldsymbol{D} \mid \boldsymbol{D}^T \boldsymbol{D} < a$*, where* $\boldsymbol{D} \sim N(0, \boldsymbol{I}_{p \times p})$.

2. *Under* $H_1 : \mu_1 - \mu_2 \neq 0$*, where* $\mu_1 - \mu_2 = \delta / \sqrt{n}$*, for a fixed* $\delta \neq 0$*; then,* $S_{\text{unadj}} \xrightarrow{d} (\delta + 2\sigma_\epsilon Z + 2\boldsymbol{\beta}^T \boldsymbol{\xi}) / \left(2\sqrt{\sigma_\epsilon^2 + \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta}} \right)$.

Furthermore, the asymptotic variance of S_{unadj} *is* $(\sigma_\epsilon^2 + m_a \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta}) / (\sigma_\epsilon^2 + \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta})$*, where* $m_a = (2\gamma(p/2+1, a/2)) / (p\gamma(p/2, a/2)) = P\left(\chi^2_{p+2} \leq a\right) / P\left(\chi^2_p \leq a\right) < 1$*, and* γ *is the incomplete gamma function* $\gamma(b, c) = \int_0^c y^{b-1} e^{-y} dy$.

Based on the corollaries, under CR, the null distribution of S_{unadj} is $N(0, 1)$. Hence, the traditional test with the critical value $z_{1-\alpha/2}$ is valid, and no adjustment is needed. Under RR, the null distribution of S_{unadj} is narrower than the standard normal distribution, indicating that the aforementioned traditional test is conservative. Here, we need to use the corrected critical value $z^{\text{RR}}_{1-\alpha/2}$ based on Corollary 3 to maintain a valid type-I error.

Because the outcome variable follows model (3.2), the covariate distributions influence the distributions of both the estimated treatment effect $\hat{\tau}$ and the test statistic S_{unadj} obtained from (3.3). Because the covariate-adaptive design changes the covariate distributions, it also distorts the dis-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

tributions of $\hat{\tau}$ and S_{unadj}. Therefore, the covariate distributions and the random nature of the covariate-adaptive design play an important role in determining the sampling distributions of $\hat{\tau}$ and S_{unadj}.

Under the ARM method, CR, and RR, the null (and alternative) distributions of S_{unadj} all share the same mean. The null (and alternative) distributions of S_{unadj} under the ARM method are the narrowest of all three, and RR is narrower than CR. When using the traditional critical value $z_{1-\alpha}$, the test under the ARM method is most conservative, with the lowest type-I error. The test under RR is moderately conservative. The test under CR is valid, with a correct type-I error. On the other hand, when using the corrected critical values, the tests are all valid. The test under the ARM method is the most powerful, and RR is more powerful than CR; see Table 2 and Figure 1.

Table 2: Comparison of ARM, CR, and RR in terms of the type-I errors of the traditional test using $z_{1-\alpha}$ and the test using the corrected critical value, and the power of the test using the corrected critical value.

Methods	Type I error of the traditional test $\lvert S_{\text{unadj}}\rvert > z_{1-\alpha/2}$	Type I error of the adjusted test $\lvert S_{\text{unadj}}\rvert >$ corrected CV	Power of the adjusted test $\lvert S_{\text{unadj}}\rvert >$ corrected CV
CR	Valid (α)	Valid	Least powerful
RR	Moderately conservative ($< \alpha$)	Valid	Moderately powerful
ARM	Most conservative ($< \alpha$)	Valid	Most powerful

For completeness, we also derive the properties of S_{adj} under ARM, CR, and RR below.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

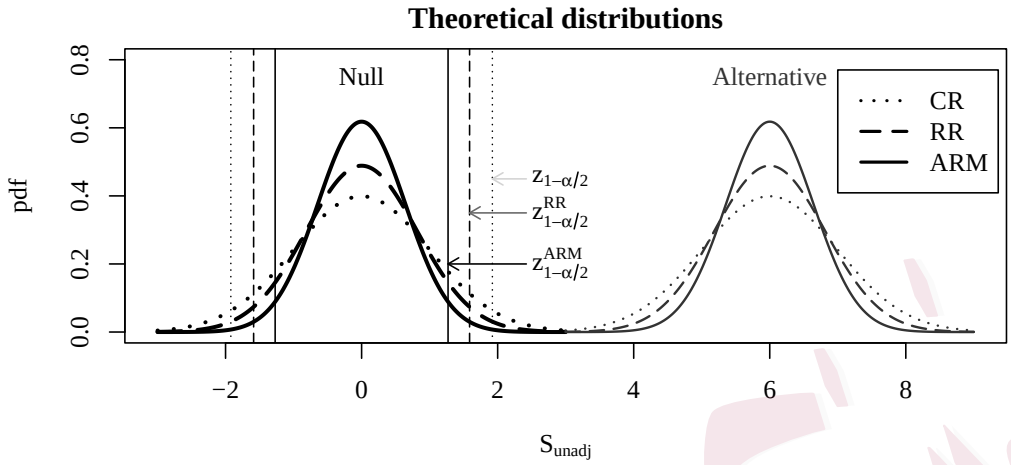


Figure 1: Comparison of the theoretical null and alternative distributions of S_{unadj} under ARM, CR, and RR.

Theorem 4. *Under ARM, CR, and RR, when testing the treatment effect using* S_{adj} *(i.e., the linear-regression-adjusted T-test), we have*

1. *Under* $H_0 : \mu_1 - \mu_2 = 0$*; then,* $S_{\text{adj}} \xrightarrow{d} N(0, 1)$.

2. *Under* $H_1 : \mu_1 - \mu_2 \neq 0$*, where* $\mu_1 - \mu_2 = \delta/\sqrt{n}$*, for a fixed* $\delta \neq 0$*; then,* $S_{\text{adj}} \xrightarrow{d} N(\delta/(2\sigma_\epsilon), 1)$.

Therefore, the null distribution of S_{adj} is the standard normal distribution, and the traditional testing procedure using the critical value $z_{1-\alpha/2}$ is valid with a correct type-I error, and no adjustment is needed. However, such an approach requires that we estimate of the working model parameters and adopt the corresponding assumptions.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

3.3 Computational Advantage

The previous section clearly demonstrates the advantages of the proposed method. A natural question is whether we can also let $a \to 0$ inRR to improve its performance to match that of the proposed method (because RR allows researchers to increase the power of the analysis at the expense of computational time (Morgan and Rubin, 2012)). However, this option is extremely expensive computationally in many cases.

Theorem 5. *For RR, to achieve the same level of covariate balance as that of the ARM method, the acceptance probability P_a of RR is $\chi^2_{df=p}(a^*)$, where $\chi^2_{df=p}(\cdot)$ is the cumulative distribution function of a chi-squared distribution with p degrees of freedom, and a^* is the root of $\gamma(p/2, a^*/2)Dp^2 = 2\gamma(p/2+1, a^*/2)n$, where $D > 0$ is a constant.*

We report the acceptance probabilities for several scenarios as quantitative values in Table 3. For a small sample size and low-dimensional covariates, the acceptance probability is reasonable. However, as either p or n increases, the acceptance probability approaches zero very quickly.

4. Numerical Studies

In this section, we use simulation studies to demonstrate the computational advantages of the proposed method. Additional numerical results are shown

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

Table 3: Acceptance probabilities of RR to match the covariate balance produced by the proposed method for different levels of n and p.

n	$p = 2$	$p = 5$	$p = 10$	$p = 20$	$p = 30$
1000	0.019360138	5.889118e-04	1.366763e-05	2.041414e-07	2.886993e-08
2000	0.009504544	1.058795e-04	4.742458e-07	3.091250e-10	2.424319e-12
3000	0.006528596	3.886533e-05	6.451756e-08	6.184287e-12	7.804135e-15

in the Supplementary Material.

4.1 Proposed Method under Different Settings

We examine the effect of the sample covariance matrix on our proposed method under three settings: (1) Causal inference setting: all units are available for assignment before the randomization starts; hence, we can use all units to estimate the covariance matrix. The sample covariance matrix stays the same throughout the randomization process. (2) Clinical trial setting: units come to the study sequentially and are assigned to a treatment sequentially; hence, we can estimate the covariance matrix using only the available units. Therefore, the sample covariance matrix needs to be updated during the randomization process. (3) Oracle setting: we use the true covariance matrix in our randomization process. We examine both the ARM and the mARM methods, and we allocate one unit at a time. Note that when allocating one unit at a time, the numbers of units in both treatment groups are usually different, with an order of $\sqrt{n}$.

We simulate the covariates according to $\boldsymbol{x}_i \sim \text{MN}(0, \boldsymbol{I}_{p\times p})$, with differ-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

ent p and n, to obtain the treatment assignments. We plot the distribution of the Mahalanobis distance of the ARM method in Figure 2, and that of the mARM method in the Supplementary Material. Blue, green, and red curves correspond to settings (1), (2), and (3), respectively. These figures show that the distributions of the Mahalanobis distance under the three settings are almost identical, especially when the sample sizes are large. Thus, the sample covariance matrix has a very limited impact on the final Mahalanobis distance. When the sample size is 200 and the number of covariates is 20, there is a mildly negative effect. This is because we need to estimate many parameters in the covariance matrix using limited observations. As more units are assigned, the sample covariance matrix converges, and the Mahalanobis distances in the three scenarios become are the same. Therefore, as long as the sample size is at least moderate, we can use the sample covariance matrix for the proposed method. This simulation study verifies the applicability of the proposed method in both causal inference and clinical trial settings with pairwise and sequential allocation procedures.

4.2 Covariate Balance and Computational Advantage

In this section, we compare the proposed method with other methods, especially RR, in terms of covariate balance and computational feasibility.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

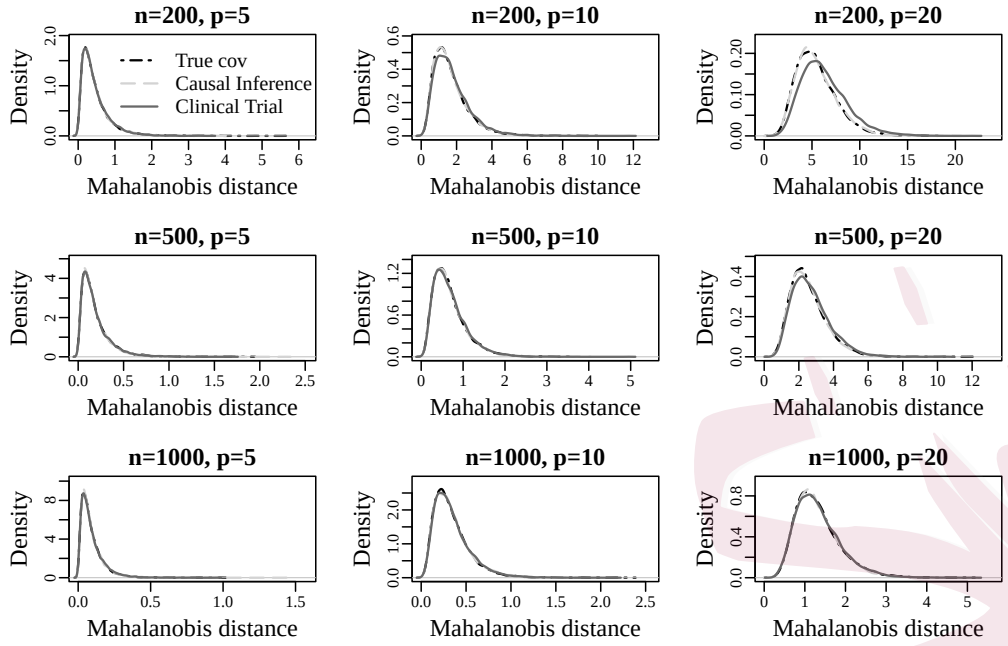


Figure 2: Comparison of the distributions of the Mahalanobis distances obtained using the ARM method under three scenarios. The short dotted curves are for known true covariance settings. Long dotted curves are for the causal inference setting. Solid curves are for the clinical trial setting.

We first compare the proposed method with RR (with $P_a = 0.05$) by simulating the covariates with $\boldsymbol{x} \sim \text{MN}(0, \boldsymbol{I}_{p\times p})$; the results are presented in Figure 3. For different n and p, we plot histograms of $M(n)$ of the proposed method and $M_{\text{RR}}(n)$ of RR. As the figure shows, as n increases, the distribution of $M_{\text{RR}}(n)$ remains unchanged, whereas the distribution of $M(n)$ converges rapidly to zero. Moreover, as p increases, the distributions obtained using RR and the proposed method become wider, but the inflation of the distribution is much less severe for the proposed method (i.e.,

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

the overlap between the two distributions becomes smaller as p increases).

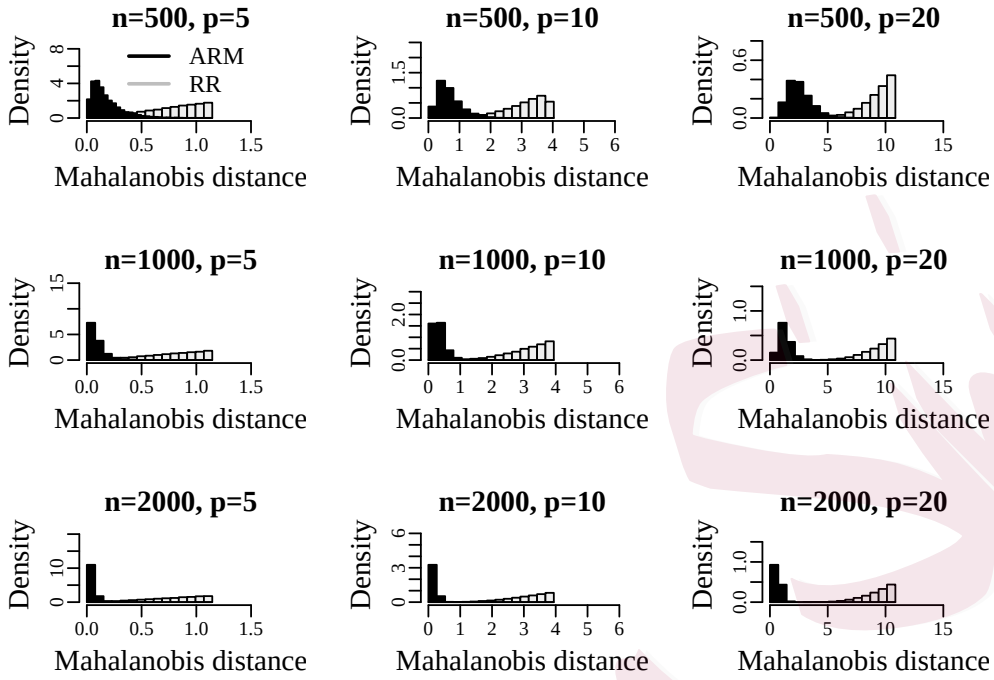
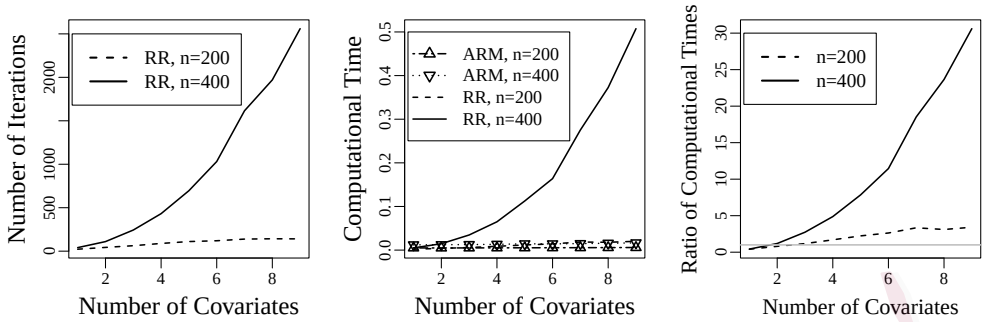


Figure 3: Comparison of the distributions of the Mahalanobis distances obtained using the proposed method, $M(n)$, and RR, $M_{\mathrm{RR}}(n)$, for different sample sizes n and different numbers of covariates p.

Next, we compare the proposed method with RR in terms of computational time. Note that the proposed method requires only one iteration, whereas RR requires multiple iterations of CR to achieve an acceptable balance level. Therefore, we compared the number of iterations required for RR to achieve the same performance (i.e., the same Mahalanobis distance) as that of the proposed method. We also compared the corresponding computational times; see Figure 4. As shown in Figures 4a and 4b, when n

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440



(a) Number of iterations (b) Computational time (c) Ratio of times

Figure 4: Comparison of the numbers of iterations, computational times, and ratios of computational times for RR and the proposed method. Panel (a): numbers of iterations of RR required to achieve the same performance as that of the proposed method. Panel (b): the corresponding computational times used in Panel (a). Panel (c): the ratios of the computational times shown in Panel (b).

and p are small, the computational advantage of the proposed method is not obvious. However, as n and p increase, the advantage of the proposed method becomes more significant, because RR requires more iterations and more time to achieve the same level of performance as that of the proposed method. As p continues to increase, RR becomes computationally expensive. Note that the computational time of the proposed method grows only linearly with n, and remains the same for different p, whereas the computational time of RR grows exponentially as either n or p increases.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

4.3 Treatment Effect Estimation

We compare the proposed method with other randomization methods in terms of the treatment effect estimation. The competing randomization methods are CR, RR, the covariate-adaptive biased coin design (CA-BCD) (Antognini and Zagoraiou, 2011), the optimum biased coin designs (DA-BCD) (Atkinson, 1982), the stratified permuted blocks randomization (SPBR) (Taves, 1974), the stratified biased coin design (SBCD) (Shao et al., 2010), ARM, and mARM. For ARM, mARM, and SBCD, the treatment allocation probability is $q = 0.75$. The acceptance probability of RR is $P_a = 0.05$. Following (Antognini and Zagoraiou, 2011), the design parameter of CA-BCD is two. The block size of SPBR is four. We simulate 10 continuous covariates $\boldsymbol{x}_i = (x_{i1}, ..., x_{i10})^T$ according to $\boldsymbol{x}_i \sim \text{MN}(\boldsymbol{0}, \boldsymbol{I}_{10\times 10})$. We set the sample size to $n = 500, 5000$ to approximate the finite/large sample size cases, respectively. We applied different randomization methods to these simulated units and obtained the simulated treatment assignments T_i. Because CA-BCD, SPBR, and SBCD cannot be used for continuous covariates, we discretize the covariates into binary covariates according to their signs. We further simulate the outcome variable according to $y_i = \mu_1 T_i + \mu_2(1 - T_i) + \sum_{j=1}^{10} \beta_j x_{ij} + \epsilon_i$, where $\mu_1 = 0$, $\mu_2 = 1$, $\beta_j = 1$, for $j = 1, ..., 10$, and $\epsilon_i \sim N(0, 2^2)$. We also compared the performance of

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

these methods in the discrete covariate case. The results are provided in the Supplementary Material.

Using the simulated data, we estimate the treatment effect using the following four working models, and obtain the asymptotic standard error for each method under the different randomization methods:

W1: $y_i = \mu_1 T_i + \mu_2(1 - T_i) + \epsilon_i$

W2: $y_i = \mu_1 T_i + \mu_2(1 - T_i) + \sum_{j=1}^{3} \beta_j x_{ij} + \epsilon_i$

W3: $y_i = \mu_1 T_i + \mu_2(1 - T_i) + \sum_{j=4}^{10} \beta_j x_{ij} + \epsilon_i$

W4: $y_i = \mu_1 T_i + \mu_2(1 - T_i) + \sum_{j=1}^{10} \beta_j x_{ij} + \epsilon_i$.

Note that W1 is equivalent to $\hat{\tau}$ and W4 is equivalent to $\tilde{\tau}$. The results are presented in Table 4, and are consistent with those shown in Table 1. Under W1, the proposed method performs best. CR performs the worst, because it does not reduce the covariate imbalance. RR and DA-BCD perform better than CR, but still worse than the proposed method. Because CA-BCD, SPBR, and SBCD are designed for discrete covariates, their performance is worse than that of the proposed method, because some covariate information is lost. When $p = 10$, the number of possible strata is 2^{10}. When the sample size is 5000, CA-BCD outperforms CR, but they perform similarly when the sample size is 500, because there are not enough observations in each stratum.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

Table 4: Comparison of the asymptotic standard errors of the estimated treatment effect for working model W1, W2, W3, and W4 under different randomization methods. This table is a verification of Table 1. Asymptotic standard errors are multiplied by $\sqrt{n}/2$ for easy comparison.

Randomization	$n = 500$, Working models				$n = 5000$, Working models			
method	W1	W2	W3	W4	W1	W2	W3	W4
ARM	2.1476	2.1145	2.0319	1.9922	2.0102	2.0043	2.0045	1.9983
mARM	2.0724	2.0425	2.0190	1.9882	2.0107	2.0093	2.0081	2.0069
CR	3.7242	3.3605	2.6168	2.0100	3.7840	3.3038	2.6565	2.0119
RR	2.7117	2.5436	2.2492	2.0465	2.6887	2.5142	2.2352	2.0075
DA-BCD	2.4212	2.3094	2.0995	1.9663	2.4483	2.3265	2.1370	2.0082
CA-BCD	3.7004	3.2656	2.6476	2.0027	3.1655	2.9052	2.3855	2.0068
SPBR	3.6204	3.2553	2.6261	1.9818	2.9297	2.6586	2.3390	1.9902
SBCD	3.5961	3.1802	2.5770	2.0072	3.1217	2.8716	2.3678	2.0088

As we include additional covariates into the working model, the standard error gradually decreases, because the covariate imbalance is partially adjusted by the linear regression. When all covariates are included in the working model (i.e., W4), the standard errors are the smallest. Note that W4 is almost the same under all randomization methods. This is because the covariate imbalance from the randomization methods has been completely adjusted, therefore; thus, the standard error reaches its minimum.

4.4 Hypothesis Testing

Here, we compare different randomization methods in terms of hypothesis testing under the same settings as those in the previous section, with $n = 5000$. For CE, RR, ARM, and mARM, we can estimate the critical values. For CA-BCD, DA-BCD, SPBR, and SBCD, we obtain the true critical

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

Table 5: type-I error for the treatment effect under various working models and randomization procedures.

Randomization method	Working models			
	W1	W2	W3	W4
ARM	0.052	0.054	0.055	0.050
mARM	0.055	0.053	0.052	0.048
CR	0.043	0.041	0.055	0.052
RR	0.037	0.042	0.050	0.048
DA-BCD	0.050	0.050	0.050	0.050
CA-BCD	0.050	0.050	0.050	0.050
SPBR	0.050	0.050	0.050	0.050
SBCD	0.050	0.050	0.050	0.050

value through simulation. We present the type-I errors in Table 5, which shows that all type-I errors with estimated critical values are successfully controlled at 5%. Therefore, the theoretical asymptotic distributions of S_{unadj} under CR, RR, ARM, and mARM work well.

We also calculate the power of the test for various levels of $\mu_1 - \mu_2$, and plot the results in Figure 5. As $\mu_1 - \mu_2$ increases away from zero, the power increases, in general. However, under different working models, the various randomization methods provide different power. ARM and mARM clearly have the highest power compared with that of the other randomization methods under the same working model. Furthermore, the power under W1 for ARM and mARM is the same as the power under W2, W3, and W4, whereas the power of the other randomization methods gradually increase from W1 to W4. Note that the different working models affect the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

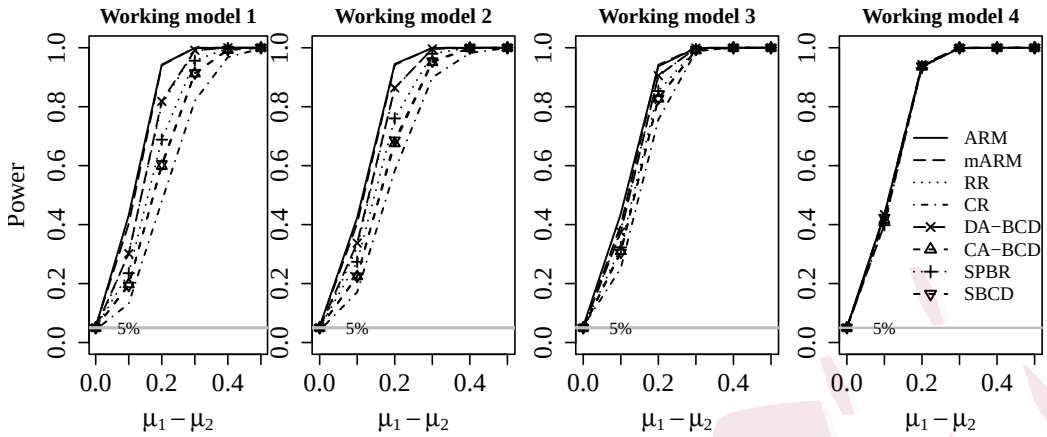


Figure 5: Power against $\mu_1 - \mu_2$. Sample size $n = 5000$. From left to right are four working models. The results for CR, RR, ARM, mARM, DA-BCD, CA-BCD, SPBR, and SBCD are shown in each panel to aid comparison.

performance of CR most, because CR does not balance the covariate. The remaining methods do balance the covariates, but not as well as ARM and mARM do, and so their power is better than that of CR but worse than that of ARM.

5. Real–Data Example

In this section, we demonstrate our proposed method using data from a real clinical study of a Ceragem massage thermal therapy bed, a device for treating lumbar disc disease. In total, there are 186 patients with $p = 50$ covariates. There are 30 continuous covariates, such as age and the baseline measurements of the patient's current conditions, for example, lower back pain and leg numbness, all measured on a scale from zero to 10. The

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

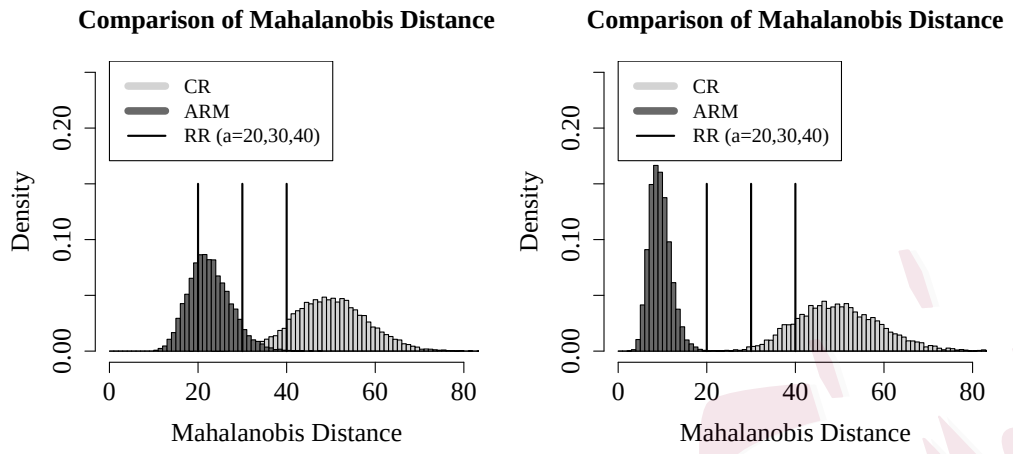
outcome variable y_i, representing measurements of lower back pain after the treatment or control experiment, was recorded to study the treatment effect.

In the original study, the patients were assigned randomly to the treatment or control groups. The corresponding Mahalanobis distance was 57.67, indicating a moderate covariate imbalance. To compare, we repeatedly assigned these patients to treatment groups using the proposed method, CR, and RR ($M < a$ and $a = 20, 30, 40$). The corresponding Mahalanobis distances are plotted in Figure 6. Note that in the right panel of Figure 6, in order to mimic a setting with a large sample, we replicated the data four times to $n = 744$.

As shown in Figure 6, the Mahalanobis distances of the proposed method on the original data ($n = 186$) are consistently lower. If we had $n = 744$ patients, the Mahalanobis distance of the proposed method decreases further toward zero. Few CR allocations achieve the same level of balance as that of the proposed method. RR produces Mahalanobis distances to the left of the vertical lines ($M = 20, 30, 40$), which are still not comparable with the proposed method.

For each randomization method, we further simulated the outcome variable y_i^{sim} according to $y_i^{\text{sim}} = \hat{\mu}_1 T_i^{\text{sim}} + \hat{\mu}_2(1 - T_i^{\text{sim}}) + \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} + \epsilon_i^{\text{sim}}$, where

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440



(a) $n = 186$ (b) $n = 744$

Figure 6: Comparison of the distributions of the Mahalanobis distance obtained using ARM, CR, and RR. Note that RR is represented by the portion of the CR distribution that lies to the left of the vertical line ($M = 20, 30, 40$).

T_i^{sim} is the simulated patient allocation, ϵ_i^{sim} is sampled from the residuals of the regression fitted to the original data, and $\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{\boldsymbol{\beta}}$ are the corresponding estimated regression coefficients. Using the simulated outcome variable, we obtain the average treatment effect using $\hat{\tau}$. The performance comparison is summarized in Table 6. The proposed method exhibits the best performance, especially when the sample size is large, and it yields the lowest variance. For RR, a smaller threshold results in better performance. However, this comes at the cost of a longer computational time and a lower acceptance probability. As the sample size increases, the gain from the proposed method becomes increasingly substantial, whereas the RR method

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

Table 6: Comparison of the proposed method with RR and CR for the real–data analysis.

Sample Size	Method	MSE (or Var)	Power of Test	
			No cov.	All cov.
$n = 186$	Proposed	0.081	0.843	0.862
	RR ($M < 40$)	0.090	0.684	0.861
	RR ($M < 30$)	0.085	0.735	0.856
	RR ($M < 20$)	0.081	0.792	0.877
	CR	0.100	0.517	0.853
$n = 744$	Proposed	0.018	0.832	0.881
	RR ($M < 40$)	0.022	0.604	0.882
	RR ($M < 30$)	0.021	0.669	0.880
	RR ($M < 20$)	0.018	0.721	0.875
	CR	0.025	0.534	0.872

does not improve at all.

We also conduct hypothesis testing for the treatment effect using the working model with no covariates and the working model with all covariates. Each test is simulated 1000 times and the results are presented in Table 6. The results show that the power of the working model that includes all covariates is similar for the various randomization procedures. However, if the working model does not include the covariates used in the randomization, the power degrades. The proposed method shows the least degradation, RR shows moderate degradation, and CR shows the most degradation. This evidence shows the importance of adjusting the hypothesis testing procedure, and that better covariate balance improves the power of the hypothesis testing. In particular, under ARM, the test has the highest power and appears equivalent to the model in which all covariates are adjusted.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

6. Discussion

We have proposed a new randomization procedure for balancing covariates in order to improve statistical inference for causal inference and clinical trials. The proposed method shows superior performance in terms of covariate balance, estimation accuracy, hypothesis testing power, and computational time. It achieves optimality under the linear regression framework, in the sense that, asymptotically, the proposed method balances the covariates so well that the imbalance adjustment provided by the linear regression is not needed.

The proposed method follows the spirit of the minimization methods used in clinical trials (Taves, 1974; Pocock and Simon, 1975; Hu and Hu, 2012). However, the focus and context of these methods differ from ours. Their methods cannot be applied when patients enroll sequentially in a clinical trial. In contrast, our proposed method can be applied both in clinical trials in which units are enrolled sequentially, and in causal inference in which all units are collected before the randomization and experiment begin. Another significant difference is that the minimization methods are suitable for discrete covariates, minimizing the margin and stratum imbalance. In contrast, the proposed method is suitable for both discrete and continuous covariates.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

Banerjee et al. (2020) and Kapelner et al. (2021) state that an optimal design should be more random than a deterministic assignment, but less random than CR. The former has the greatest robustness, but its estimation efficiency is unsatisfactory. The latter is less robust, because unseen subject-specific characteristics can be highly imbalanced, which is considered risky and leads to a severe loss in robustness. Thus, the proposed method is a good choice for a design, because it can make the trade-off between CR and the highly optimized deterministic assignment by using the treatment allocation probability q. For the proposed method, we suggest a mild q (such as $q = 0.75$) to avoid pursuing the covariate balance to the extreme.

Many directions for further research remain. For example, as the number of covariates increases, it is more efficient to balance only the most important covariates (Morgan and Rubin, 2015); therefore, selecting the important covariates to balance in our proposed framework is an interesting topic for future work.

Acknowledgments

The authors would like to thank the editor, associate editor, and reviewers for their constructive comments, which lead to a significant improvement of this work. Wei Ma's work was supported by the National Key R&D

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

Program of China (Grant No. 2018YFC2000302).

Supplementary Material

The online Supplementary Material contains additional numerical studies and proofs.

References

Antognini, A. B. and M. Zagoraiou (2011). The covariate-adaptive biased coin design for balancing clinical trials in the presence of prognostic factors. *Biometrika 98*(3), 519–535.

Atkinson, A. C. (1982). Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika 69*, 61–67.

Banerjee, A. V., S. Chassang, S. Montero, and E. Snowberg (2020). A theory of experimenters: Robustness, randomization, and balance. *American Economic Review 110*(4), 1206–30.

Bruhn, M. and D. McKenzie (2008). In pursuit of balance: Randomization in practice in development field experiments. *World Bank Policy Research Working Papers*.

Ciolino, J., W. Zhao, R. Martin, and Y. Palesch (2011). Quantifying the cost in power of ignoring continuous covariate imbalances in clinical trial randomization. *Contemporary Clinical Trials 32*(2), 250–259.

Cochran, W. G. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society, Series A 128*(2), 234–266.

Cochran, W. G. and D. B. Rubin (1973). Controlling bias in observational studies: a review. *Sankhya, A 35*(4), 417–446.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain 33*, 503–513.

Frane, J. W. (1998). A method of biased coin randomization, its implementation, and its validation. *Drug Information Journal 32*, 423–432.

Freedman, D. A. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics 40*(2), 180–193.

Hoehler, F. K. (1987). Balancing allocation of subjects in biomedical research: a minimization strategy based on ranks. *Computers and Biomedical Research 20*, 209–213.

Hu, F., Y. Hu, Z. Ma, and W. F. Rosenberger (2014). Adaptive randomization for balancing over covariates. *Wiley Interdisciplinary Reviews: Computational Statistics 6*(4), 288–303.

Hu, Y. and F. Hu (2012). Asymptotic properties of covariate-adaptive randomization. *Annals of Statistics 40(3)*, 1794–1815.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0440

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Kapelner, A., A. M. Krieger, M. Sklar, U. Shalit, and D. Azriel (2021). Harmonizing optimized designs with classic randomization in experiments. *The American Statistician 75*(2), 195–206.

Li, X. and P. Ding (2020). Rerandomization and regression adjustment. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*(1), 241–268.

Li, X., P. Ding, and D. B. Rubin (2018). Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences 115*(37), 9157–9162.

Lin, Y. and Z. Su (2012). Balancing continuous and categorical baseline covariates in sequential clinical trials using the area between empirical cumulative distribution functions. *Statistics in Medicine 31*, 1961–1971.

Lock, K. F. (2011). Rerandomization to improve covariate balance in randomized experiments. *Ph.D. Thesis, Harvard University,*.

Ma, W., Y. Qin, Y. Li, and F. Hu (2019). Statistical inference for covariate-adaptive randomization procedures. *Journal of the American Statistical Association*.

Ma, Z. and F. Hu (2013). Balancing continuous covariates based on kernel densities. *Contemporary Clinical Trials 34*(2), 262–269.

Morgan, K. L. and D. B. Rubin (2012). Rerandomization to improve covariate balance in experiments. *Annals of Statistics 40(2)*, 1263–1282.

Morgan, K. L. and D. B. Rubin (2015). Rerandomization to balance tiers of covariates. *Journal of the American Statistical Association 110*(512), 1412–1421.

Pocock, S. J. and R. Simon (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics 31*(1), 103–115.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics 2*(3), 808–840.

Scott, N. W., G. C. McPherson, C. R. Ramsay, and M. K. Campbell (2002). The method of minimization for allocation to clinical trials: a review. *Controlled Clinical Trials 23*(6), 662–674.

Shao, J., X. Yu, and B. Zhong (2010). A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika 97*(2), 347–360.

Stigsby, B. and D. R. Taves (2010). Rank-minimization for balanced assignment of subjects in clinical trials. *Contemporary Clinical Trials 31*(2), 147–150.

Taves, D. R. (1974). Minimization: A new method of assigning patients to treatment and control groups. *Clinical Pharmacology & Therapeutics 15*(5), 443–453.

Zhou, Q., P. A. Ernst, K. L. Morgan, D. B. Rubin, and A. Zhang (2018). Sequential rerandomization. *Biometrika 105*(3), 745–752.